Day 9: Unsupervised learning and dimensional reduction

Kenneth Benoit and Slava Mikhaylov

Introduction to Data Science and Big Data Analytics

27 August 2015

Day 9 Outline

Decision Trees

Types of dimensional reduction models

Dimensional reduction methods

Exam (p)Review

Additional resources for data science

Decision trees

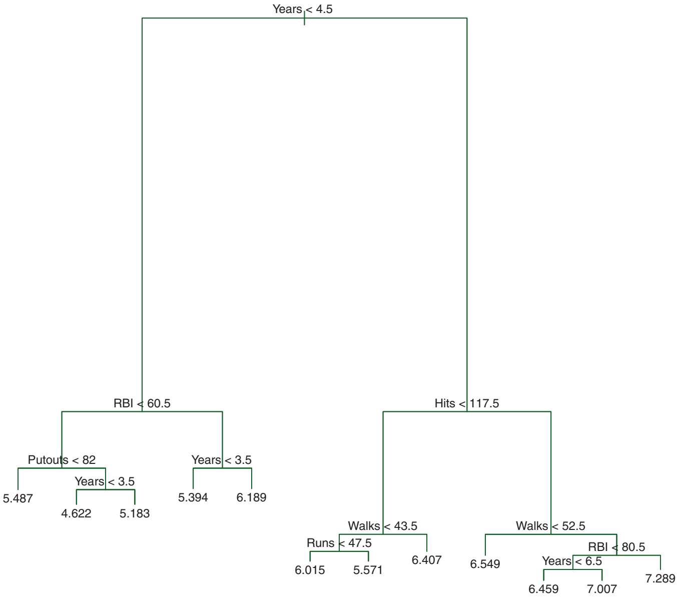
Brief Introduction to Decision Trees

- Basic idea: segment the feature space to produce a set of terminal nodes associated with an outcome, which can be used to locate future observations to produce a prediction
- Can be used for classification, if outcome is categorical rather than continuous
- Basic methods not as effective as other methods, but can be augmented with advanced versions, such as bagging, random forests, and boosting

Decision Trees: Terminology and concepts

- terminal nodes: contain mean response of predicted variable following the partition along branches of the tree
- internal nodes: points along the tree where the predictor space is split
- prediction (and construction of the trees) proceeds through stratification of the predictor space. This is done using a variety of algorithms, selecting variables in a (usually) *top-down* approach based on some variance or entropy criterion
- pruning: tree size is reduced following the algorithmic construction of a complete tree, because complete trees tend to overfit the data and lead to poor test set performance

Walk-through



Walk-through

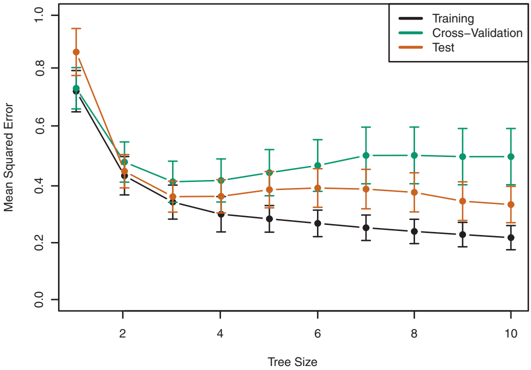
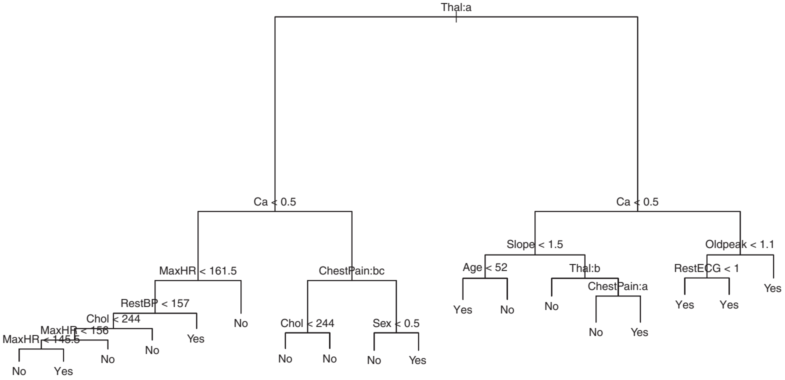


FIGURE 8.5. *Regression tree analysis for the* `Hitters` *data. The training, cross-validation, and test MSE are shown as a function of the number of terminal nodes in the pruned tree. Standard error bands are displayed. The minimum cross-validation error occurs at a tree size of three.*

Class prediction



Class prediction after pruning

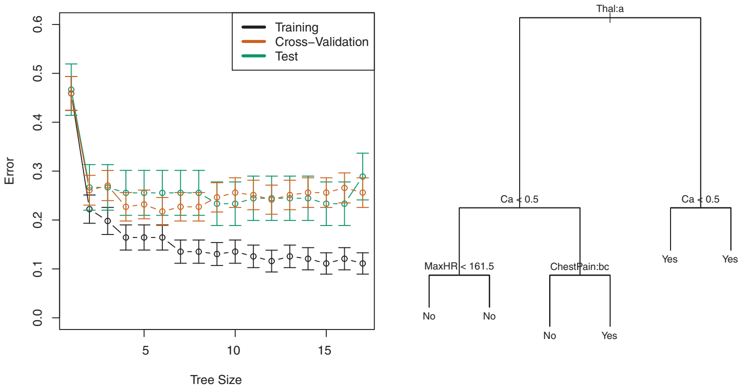


FIGURE 8.6. Heart *data.* Top: *The unpruned tree.* Bottom Left: *Cross -validation error, training, and test error, for different sizes of the pruned tree.* Bottom Right: *The pruned tree corresponding to the minimal cross-validation error.*

Types of dimensional reduction models

Parametric v. non-parametric methods

- Parametric methods model feature occurrence according to some stochastic distribution, typically in the form of a measurement model
  - for instance, model words as a multi-level Bernoulli distribution, or a Poisson distribution
  - feature effects and "positional" effects are unobserved parameters to be estimated

- Non-parametric methods typically based on the Singular Value Decomposition of a matrix
  - principal components analysis
  - correspondence analysis
  - other (multi)dimensional scaling methods

Dimensional reduction methods

Non-parametric dimensional reduction methods

- Non-parametric methods are algorithmic, involving no "parameters" in the procedure that are estimated
- Hence there is no uncertainty accounting given distributional theory
- Advantage: don't have to make assumptions
- Disadvantages:
  - cannot leverage probability conclusions given distribtional assumptions and statistical theory
  - results highly fit to the data
  - not really assumption-free (if we are honest)

Principal Components Analysis

- For a set of features $X_1, X_2, \ldots, X_p$, typically centred (to have mean 0)
- the first principal component is the normalized linear combination of the features

$$Z_1 = \phi_{11}X_1 + \phi_{21}X_2 + \ldots + \phi_{p1}X_p$$

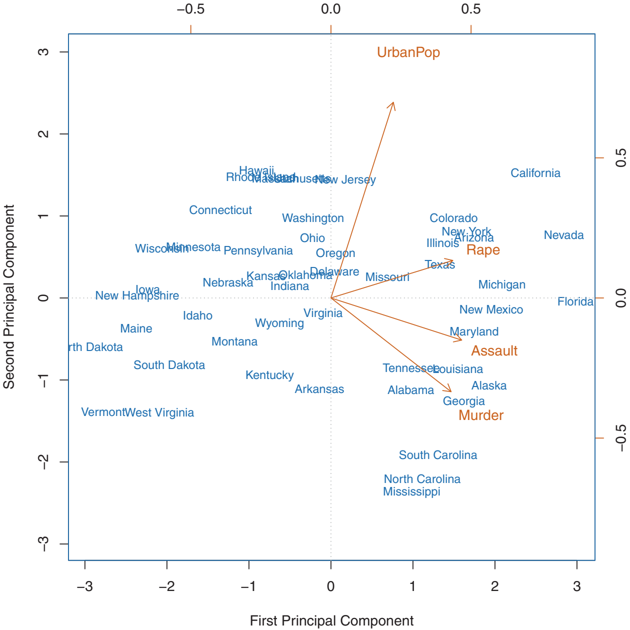
  that has the largest variance
- normalized means that $\sum_{j=1}^{p} \phi_{j1}^2 = 1$
- the elements $\phi_{11}, \ldots, \phi_{p1}$ are the loadings of the first principal component
- the second principal component is the linear combination Z_2 of $X_1, X_2, \ldots, X_p$ that has maximal variance out of all linear combinations that are *uncorrelated* with Z_1

PCA factor loadings example

	PC1	PC2
Murder	0.5358995	−0.4181809
Assault	0.5831836	−0.1879856
UrbanPop	0.2781909	0.8728062
Rape	0.5434321	0.1673186

TABLE 10.1. *The principal component loading vectors, ϕ_1 and ϕ_2, for the* `USArrests` *data. These are also displayed in Figure 10.1.*

PCA factor loadings biplot



PCA projection illustrated

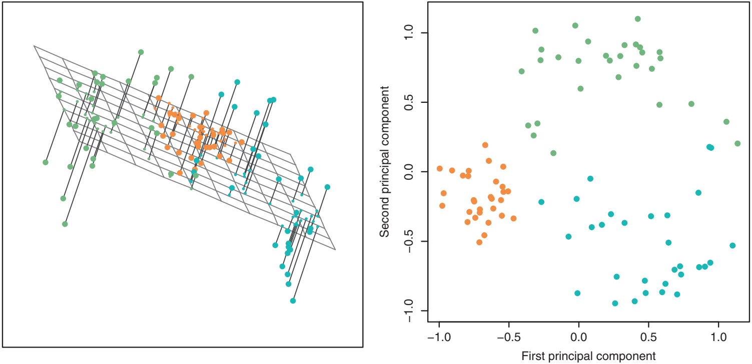


FIGURE 10.2. *Ninety observations simulated in three dimensions.* Left: *the first two principal component directions span the plane that best fits the data. It minimizes the sum of squared distances from each point to the plane.* Right: *the first two principal component score vectors give the coordinates of the projection of the 90 observations onto the plane. The variance in the plane is maximized.*

PCA projection illustrated

Table 5.6 Dimensional analysis of the Dutch policy space: principal components factor analysis, $n = 77$, parameters = 24

Factor	*Eigenvalue*	*Proportion*	*Cumulative*
1	4.28	0.48	0.48
2	1.50	0.17	0.64
3	1.26	0.14	0.78
4	0.63	0.07	0.85
5	0.52	0.06	0.91
6	0.28	0.03	0.94
7	0.26	0.03	0.97
8	0.18	0.02	0.99
9	0.08	0.01	1.00

Varimax rotated factor loadings

Variable	*Factor*		
	1 Economic Left–right	*2 EU*	*3 Social Liberalism*
Taxes vs spending	**0.88**	–0.17	0.18
Environment	**0.89**	0.15	0.01
Immigration	**0.87**	0.23	0.15
Deregulation	**0.95**	–0.09	0.07
EU accountability	**0.67**	**0.54**	0.23
EU security	–0.23	**0.84**	–0.01
EU authority	0.43	**0.71**	–0.01
Social liberalism	0.08	0.11	**0.84**
Decentralization	–0.18	0.08	**–0.80**

Correspondence Analysis

- CA is like factor analysis for categorical data
- Following normalization of the marginals, it uses Singular Value Decomposition to reduce the dimensionality of the word-by-text matrix
- This allows projection of the positioning of the words as well as the texts into multi-dimensional space
- The number of dimensions – as in factor analysis – can be decided based on the eigenvalues from the SVD

Correspondence Analysis

- CA is like factor analysis for categorical data
- Following normalization of the marginals, it uses Singular Value Decomposition to reduce the dimensionality of the word-by-text matrix
- This allows projection of the positioning of the words as well as the texts into multi-dimensional space
- The number of dimensions – as in factor analysis – can be decided based on the eigenvalues from the SVD

Singular Value Decomposition

- A matrix $\underset{i \times j}{\mathbf{X}}$ can be represented in a dimensionality equal to its rank k as:

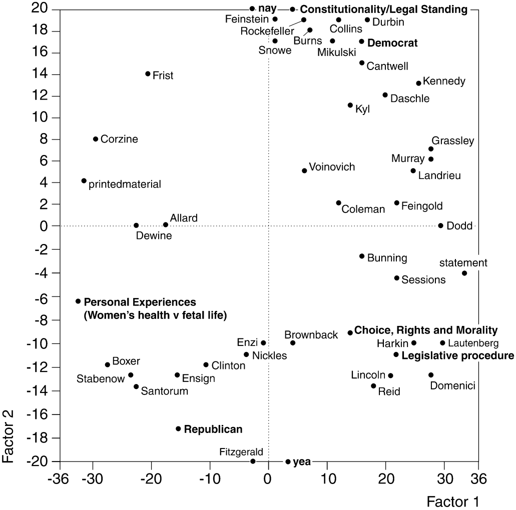
$$\underset{i \times j}{\mathbf{X}} = \underset{i \times k}{\mathbf{U}} \; \underset{k \times k}{\mathbf{d}} \; \underset{j \times k}{\mathbf{V}'} \tag{1}$$

- The **U**, **d**, and **V** matrixes "relocate" the elements of **X** onto new coordinate vectors in n-dimensional Euclidean space
- Row variables of **X** become points on the **U** column coordinates, and the column variables of **X** become points on the **V** column coordinates
- The coordinate vectors are perpendicular (*orthogonal*) to each other and are normalized to unit length

Correspondence Analysis and SVD

- Divide each value of $\mathbf{X}$ by the geometric mean of the corresponding marginal totals (square root of the product of row and column totals for each cell)
  - Conceptually similar to subtracting out the χ^2 expected cell values from the observed cell values
- Perform an SVD on this transformed matrix
  - This yields singular values $\mathbf{d}$ (with first always 1.0)
- Rescale the row ($\mathbf{U}$) and column ($\mathbf{V}$) vectors to obtain canonical scores (rescaled as $U_i\sqrt{f_{..}/f_{i.}}$ and $V_j\sqrt{f_{..}/f_{j.}}$)

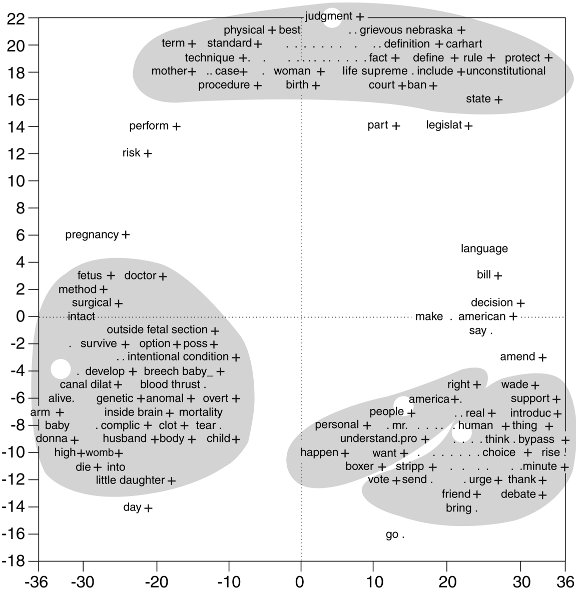
Example: Schonhardt-Bailey (2008) - speakers



	Eigenvalue	% Association	% Cumulative
Factor 1	0.30	44.4	44.4
Factor 2	0.22	32.9	77.3

Fig. 3. Correspondence analysis of classes and tags from Senate debates on Partial-Birth Abortion Ban Act

Example: Schonhardt-Bailey (2008) - words



How to get confidence intervals for CA

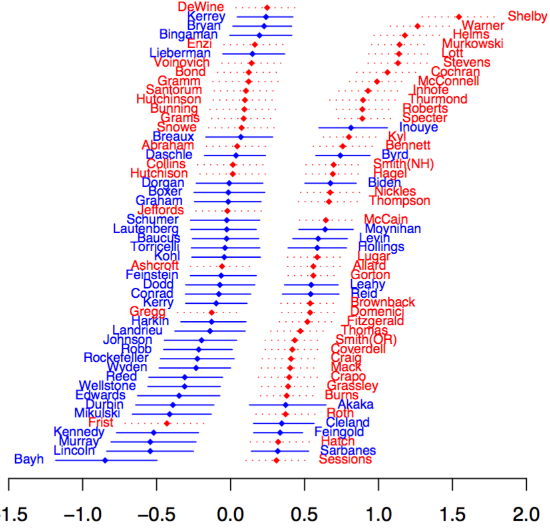
- There are problems with bootstrapping: (Milan and Whittaker 2004)
  - rotation of the principal components
  - inversion of singular values
  - reflection in an axis

How to account for uncertainty

- Ignore the problem and hope it will go away
  - SVD-based methods (e.g. correspondence analysis) typically do not present errors
  - and traditionally, point estimates based on other methods have not either

Plotting θ

Plotting θ (the ideal points) gives estimated positions. Here is Monroe and Maeda's (essentially identical) model of legislator positions:

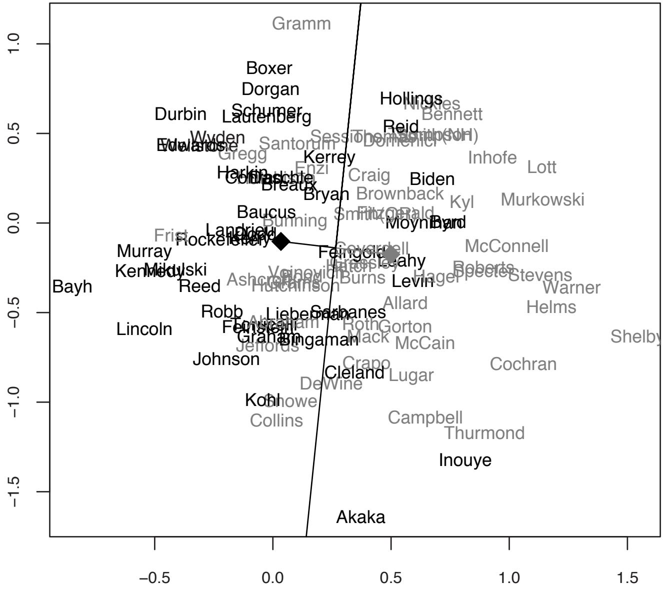


Dimensions

How infer more than one dimension?
This is two questions:

- How to get two dimensions (for all policy areas) at the same time?
- How to get one dimension for each policy area?

How do we interpret multiple dimensions?



Exam (p)Review

Exam hints

- Structure
- Topic scope
- Format

Additional resources

Additional resources

- R
  - CRAN
  - R-bloggers
  - Stack Overflow R tag
- Data science
  - lots of on-line courses
  - lots of blogs (see `http://www.ngdata.com/top-data-science-resources/`)
  - more statistics